

# Ferramentas e Tecnologias para a Integração e Extração de Informação Hospitalar

Matheus Coppetti Silveira, Vitor Beires Nogueira, and Irene Pimenta Rodrigues

Universidade de Évora  
[matheuscoppetti@gmail.com](mailto:matheuscoppetti@gmail.com)  
{vbn,ipr}@uevora.pt  
<http://www.uevora.pt>

**Resumo** Neste trabalho, apresentam-se algumas técnicas e tecnologias da *web* semântica para demonstrar a sua aplicação no campo da medicina, mais precisamente em sistemas para a identificação de informação médica em relatórios clínicos. Descrevemos a utilização da *web* semântica no desenvolvimento desses sistemas. O trabalho desenvolvido define os processos que, pela avaliação dos históricos clínicos de pacientes, identifiquem os termos de importância semântica para a recuperação da informação clínica nestes relatórios.

## 1 Introdução

Ontologias têm sido utilizadas para a representação de informações que veiculem um entendimento semântico comum de situações variadas do mundo real. Na Web, o uso de ontologias pode fornecer uma base de informações comum, bem como padronizada, englobando conceitos-chave que possam ser utilizados por serviços requisitados para cada situação particular.

As novas tecnologias têm criado novos tópicos da investigação relacionada com as biociências. Nomeadamente, o desenvolvimento de sistemas de informação 'inteligentes' no domínio da biologia e da medicina revelou a importância destes na administração e gestão do conhecimento médico [15].

A *web* semântica permite criar metadados acessíveis automaticamente a máquinas. O software e as metodologias já desenvolvidas podem ser aplicados a domínios específicos para serem capazes de fazer o uso extensivo de especificação em uma linguagem expressiva, que se estende a ações que podem definir a terminologia de um domínio científico como uma ontologia que pode ser interpretada por uma máquina. Estas ontologias foram desenvolvidas e melhoradas ao longo do avanço da Web Semântica, podem ser exploradas para a representação de conhecimento, que pode ser usado automaticamente por máquinas e podem incluir informação específica da área da medicina, onde a homogeneidade da terminologia é particularmente problemática.

As ontologias têm sido desenvolvidas para facilitar a partilha de conhecimento. São pilares das tecnologias semânticas, pois fornecem vocabulários estruturados que descrevem uma especificação formal de um conceito compartilhado

[9], possibilitando às máquinas aceder a uma semântica formal para apoiar o seu raciocínio. Várias abordagens para a Inteligência Artificial têm trabalhado em ambientes complexos como a área médica. Estas tecnologias semânticas podem fornecer uma base consistente para os sistemas médicos orientados para o conhecimento.

Dadas as exigências e os requisitos tecnológicos que são necessários para o desenvolvimento e desempenho de sistemas para a identificação semântica em textos médicos, é necessário rever uma ampla gama de tecnologias que os possam reforçar e ofereçam qualidade à concepção e implementação desses sistemas. Este trabalho tem como objetivo o desenvolvimento de um sistema para a identificação de termos em narrativas médicas sob a forma de texto livre e ainda, permitir a extração destas informações em bases de dados heterogêneas, fazendo o uso de técnicas e ferramentas de integração de informação hospitalar.

## 2 Background

Para ser possível a compreensão das tarefas executadas na identificação de termos médicos e na população de ontologias, bem como a recolha de dados médicos em bases de dados heterogêneas, é necessário um entendimento em diferentes tecnologias que se aplicam à informática médica, sendo elas a Integração de Informação Hospitalar, o Processamento de Língua Natural e a Web Semântica.

### 2.1 Integração de Informação Hospitalar

A área da informática clínica sofreu uma grande expansão nas últimas décadas. Inicialmente, apenas existiam demonstrações simples de vários processos baseados na tecnologia de informação. Com o passar do tempo, a informática clínica tem ganho cada vez mais importância, gerando investimento e aumentando a sua aplicação.

À medida que essas tecnologias foram se tornando mais disponíveis, também aumentou a necessidade de avaliações de alta qualidade para conceder uma eficiência científica. Com as pesquisas para comparação de eficácia [12], houve um crescimento na necessidade de desenvolver novos métodos para realizar avaliações mais rigorosas dos aspectos na tecnologia de informação para a saúde<sup>1</sup>[22].

Um dos campos da HIT que tem ganho relevância é Health Information Exchange (HIE), que pretende partilhar informações clínicas entre diferentes instituições [24]. Como prova da importância desta área na HIT, podemos constatar o elevado investimento realizado pelo Office of the National Coordinator for Health Information Technology (ONC) para pesquisas em HIE [25].

#### 2.1.1 Health Level 7

O Health Level 7 (HL7) é um padrão para troca de mensagens (clínicas, administrativas ou financeiras) entre diferentes sistemas de informação hospitalar

---

<sup>1</sup> Health Information technology (HIT)

(Schloeffel et al., 2006). Esse padrão define as transações necessárias para transmitir dados sob um formato que seja reconhecido por diversos sistemas [6]. O HL7 não é um software ou uma ferramenta, mas sim um conjunto de regras para enviar mensagens através de blocos de texto que podem ser transferidos entre diversos sistemas [7]. A versão 2 deste padrão teve uma grande aceitação mundial nos sistemas de informação hospitalar, com uma taxa de adoção de 95% nos Estados Unidos. Actualmente, este padrão encontra-se na terceira versão, lançada em 2005, que apesar de solucionar problemas da versão anterior, não obteve uma grande aceitação global, principalmente pela falta de compatibilidade com as versões existentes, pela estrutura complexa e fraca documentação [7].

## 2.2 Processamento de Língua Natural

Processamento de Linguagem Natural (PLN) é definido como um conjunto de técnicas computacionais, que podem tanto ser de hardware ou software de análise ou síntese de linguagem escrita ou falada. O PLN escrito envolve o uso de conhecimentos léxicos, sintáticos, e semânticos do idioma bem como outras informações reais necessárias. Oferece para a pesquisa médica, uma possibilidade de, segundo Hripksac "libertar o conteúdo clínico de relatórios narrativos".

O apoio à decisão clínica computadorizada, *Clinical Decision Support* (CDS) tem como objetivo auxiliar a tomada de decisão dos profissionais de saúde e ao público, fornecendo informações relacionadas à saúde de fácil acesso, no momento e no tempo que for necessário. O PLN é fundamental na utilização de informações em texto livre para auxiliar os CDS, aproveitando a narrativa clínica para representar intervenções de conhecimento clínico e suporte à decisão. As pesquisas em PLN foram caracterizadas por um período de investigação estável realizadas nos principais centros clínicos e de uma mudança de foco para PLN biomédica, concentrando-se principalmente no interesse renovado recentemente pelo desenvolvimento de métodos e avanços fundamentais da PLN nos sistemas de CDS. É discutido aqui o Reconhecimento de Entidades Nomeadas<sup>2</sup> como uma abordagem do processamento de língua natural (PLN) para o campo de estudo da medicina.

### 2.2.1 Named Entity Recognition

O reconhecimento de entidades nomeadas (NER) é um importante ramo no PLN, tem como tarefa a classificação e busca de expressões em textos de língua natural. Estas expressões vão desde nomes de pessoas e organizações a datas, possuindo dessa forma informações de valor sobre o texto.

O NER pode ser utilizado em diversas actividades de PLN, como por exemplo, uma ferramenta para busca e filtro em um texto. Também, pode ser utilizado para realizar o pré-processamento para outras tarefas de PLN. As tarefas que utilizam NER são a tradução automática, sistemas de pergunta-resposta, sumários, modelação de língua natural e análise de sentimento em textos, e é

<sup>2</sup> Named Entity Recognition (NER)

retirando as vantagens do uso de entidades nomeadas (NE) e trabalhando com elas individualmente, que essas tarefas obtém melhores performances.

O NER foi primeiramente introduzido no MUC-6 em 1995, e desde então os sistemas para NER estão migrado de aplicações baseadas em regras para modelos estatísticos. O estado da arte para a língua inglesa encontra-se nos 90%, sendo assim, necessário preencher esta lacuna de performance em outros idiomas.

### 2.3 Ontologias

A definição num dicionário clássico para o termo ontologia, identifica-o como "o ramo da metafísica que estuda a natureza da existência". Em aplicações do mundo real, porém, ontologia é um tipo de entidade computacional e não deve ser considerado como uma entidade natural, mas sim, como um recurso artificialmente criado [16]. Uma ontologia deve ser compreendida como um entendimento partilhado e comum sobre um determinado domínio [10].

De acordo com a definição dada por Gruber [10], uma ontologia é "uma especificação formal, explícita de uma conceitualização compartilhada". A palavra "explícita" refere-se à necessidade de especificar conscientemente os vários itens que compõem uma ontologia, a especificação formal deve ser representada numa linguagem formal. "Compartilhada" significa que a ontologia deve, na melhor das hipóteses, ter em conta o conhecimento aceito (pelo menos para o grupo de pessoas a usá-lo).

Quando se fala de ontologias como "sistemas de representação do conhecimento", deve-se especificar o tipo de sistema a que nos referimos. Na realidade, as ontologias são usadas em todos os tipos de aplicações em que é necessário definir com precisão o conjunto de entidades relevantes de uma área específica de aplicação e as interações entre elas. Algumas ontologias são criadas com o simples objetivo de se chegar a um entendimento do universo do discurso (Universe of Discourse - UoD), pois a sua criação exige uma especificação muito detalhada. Outras ontologias foram criadas com um propósito geral, como o projeto Cyc [11] que visa a construção de uma base de conhecimento humano para fazer inferências.

## 3 Metodologia

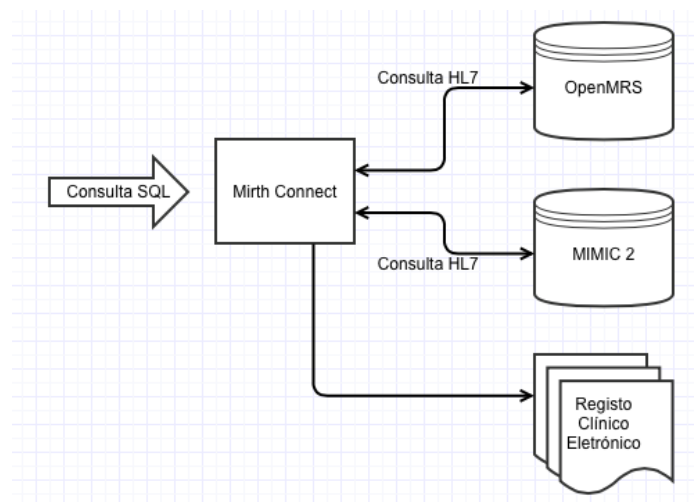
Com base nas tecnologias descritas na secção anterior foi possível o desenvolvimento de uma ferramenta capaz de realizar a integração da informação hospitalar e ainda, identificar os termos médicos presentes nestes relatórios para, assim, popular ontologias de domínio clínico. Nesta secção serão detalhadas essas tarefas.

### 3.1 Obtenção de informação em bases de dados heterogêneas

Foram utilizadas duas bases de dados com informações de relatórios clínicos de pacientes, sendo essas especificamente a base de dados do OpenMRS [26] e a base de dados MIMIC 2 [19].

Uma vez que as bases de dados utilizadas possuem uma estrutura heterogênea foi necessária a utilização de uma ferramenta que possibilitasse a integração da informação contida nessas. Como opção para tal tarefa foi utilizado o Mirth Connect [3], que permitiu através de consultas sob as normas HL7 efetuar a recolha das informações contidas nas duas bases de dados.

Ainda, relativamente às bases de dados, para manter uma maior fidelidade nas informações recolhidas, não foram feitas entradas de dados duplicadas, e foram cadastrados os mesmos pacientes em cada uma delas, porém, com relatórios médicos diferentes para esses. Desta forma, pôde-se simular o comportamento de uma ferramenta de recolha de informações em diferentes sistemas hospitalares, permitindo analisar o histórico clínico de um paciente de uma forma mais ampla, não se restringido a apenas um sistema hospitalar.



**Figura 1.** Integração da informação nas bases de dados MIMIC 2 e OpenMRS

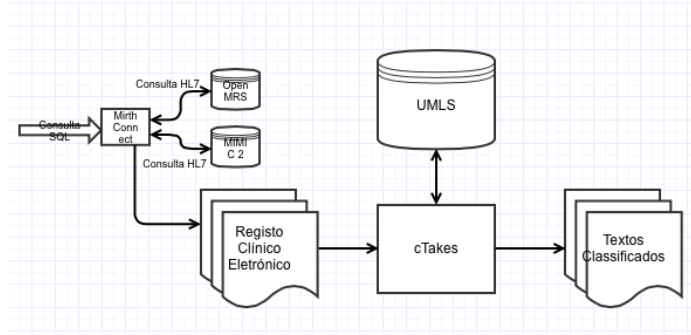
A figura 3.1 mostra um diagrama desta integração de informação hospitalar. O diagrama mostra que, a partir de uma entrada de dados sob a forma de uma consulta SQL, o Mirth Connect vai mapear esta consulta para cada uma das bases que se deseja consultar. Este mapeamento é feito pelo utilizador que, ao criar um canal<sup>3</sup> para comunicação com a base de dados, e recebendo através deste mesmo *channel* as informações clínicas do paciente. Tendo sido obtido o registo clínico do paciente, o Mirth Connect estrutura esta informação em arquivos.

### 3.2 Identificação dos termos clínicos

Após a recolha dos dados na etapa de integração de informação hospitalar, os arquivos gerados pelo Mirth Connect foram analisados para assim realizar a

<sup>3</sup> o *Mirth* trata um canal em seu sistema pelo termo em inglês *channel*

classificação destes. A classificação feita foi tanto de carácter linguístico (classificação léxica) como também de interesse clínico, sendo assim identificados os termos médicos correspondentes a actos médicos, sintomas, entidades anatómicas e doenças. Tal classificação foi feita utilizando a ferramenta considerada como o estado da arte no processamento de língua na área médica, o cTakes [20].



**Figura 2.** Processamento dos arquivos pelo cTakes

Na figura 3.2 é mostrada a etapa de classificação e identificação de termos médicos nos textos gerados pelo Mirth Connect. Os arquivos na sintaxe HL7 são importados pelo cTakes e, através de um *pipeline* são processados de modo que seja feita a classificação léxica dos textos e, com o uso de recursos UMLS disponibilizados pela ferramenta, são identificados os termos médicos nos relatórios. Esses termos são ainda classificados sob duas formas, sendo a primeira delas o *medfacts* que vai classificar o termo em PROBLEM (uma condição patológica), ANATOMICAL\_SITE (parte do corpo) e ACT (procedimento médico). Ainda, em uma segunda classificação, estes *medfacts* são classificados de acordo com seus códigos na terminologia clínica SNOMED CT [8], sendo que para cada *medfact* pode haver uma ou mais identificações na terminologia citada.

Após feita a classificação dos textos, o cTakes gera como saída uma série de arquivos XML. Tais arquivos são tratados e utilizados para fins semânticos como pode ser visto a seguir.

### 3.3 População de ontologias médicas

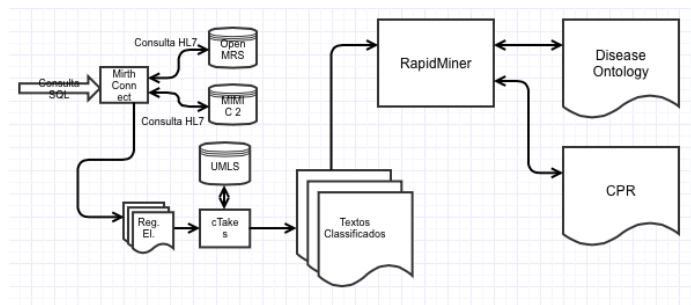
Nesta etapa, os arquivos XML contendo os textos classificados pelo cTakes passam por uma série de tratamento para que assim, possam ser extraídas as informações desejadas para realizar a população de duas ontologias de domínio clínico. Sobre os termos classificados, foram considerados para esta tarefa apenas aqueles identificados nos *medfacts* como PROBLEM, desta forma, o sistema trata apenas de condições patológicas identificadas no registo clínico de um paciente.

As ontologias escolhidas atendem às necessidades deste trabalho e são mantidas pela OBO Foundry, estando assim de acordo com as normas de qualidade e padronização exigidas por essa comunidade [21]. Foram então escolhidas duas ontologias, a *Computer-based Patient Record (CPR)* e a *Disease Ontology (DO)*.

A CPR é capaz de representar grande parte do conhecimento presente nos actos médicos e nos registos eletrônicos dos pacientes, os *electronic health records (EHRs)*, como os dados pessoais do paciente e sua condição patológica, além de diversas outras informações que fazem parte de um EHR.

A *Disease Ontology* por sua vez representa o conhecimento sobre as doenças, bem como a sua natureza, descrição, códigos SNOMED CT, ICD9 e as suas relações hierárquicas.

Utilizando as duas referidas ontologias é possível então obter o conhecimento do histórico clínico de um paciente através da CPR e, extrair o conhecimento de sua situação patológica com a semântica representada na DO. A ligação entre essas duas ontologias é feita através de uma propriedade *is\_a*, que define que uma *disease* (e suas subclasses) na DO é uma *pathological\_condition* na CPR. Ainda, na CPR, uma *pathological\_condition* possui uma propriedade *located\_in* que vai definir onde esta situação patológica ocorre, no caso, em qual o paciente ela foi identificada.



**Figura 3.** Processos para a identificação dos termos a serem utilizados para popular as ontologias

Pode ser visto na figura 3.3 como se dá a relação entre os arquivos gerados pelo cTakes, a ferramenta de mineração de dados RapidMiner e as ontologias CPR e DO. O RapidMiner atua como uma ferramenta de mineração de dados [14], onde serão extraídas as informações de interesse nos arquivos XML gerados pelo cTakes através de um filtro *xPath* para selecionar os *medfacts* marcados como PROBLEM. Com base no *id* destes registos retornados pelo filtro, é utilizado um módulo disponibilizado no RapidMiner para a escrita de códigos Java para estender a sua utilização. Neste módulo, foi escrito um *script* que vai selecionar, com base nos identificadores dos *medfacts* encontrados, quais as entradas SNOMED CT para cada um destes registos.

O RapidMiner através do módulo RMonto [17] carrega as duas ontologias para que, com base nos registos SNOMED CT identificados e com o uso de outro módulo Java seja possível identificar quais as classes de interesse nas ontologias para assim criar as instâncias e as relações entre elas para a representação do conhecimento clínico no registo clínico do paciente.

Identificar as classes de interesse na ontologia DO é uma tarefa que exige consultas sobre a ontologia de modo a identificar as classes que possuem a propriedade *database\_cross\_reference* com o código SNOMED desejado. É com a identificação da classe que possui este código SNOMED que é possível criar uma instância para a condição patológica, neste caso uma doença, identificada nos registos do paciente.

Sob um ponto de vista mais prático, pode ser tomada como entrada um arquivo XML como o descrito no trecho de código abaixo.

```
<CAS version="2">
  <uima.cas.Sofa _indexed="0" _id="3" sofaNum="1"
    sofaID="_InitialView" mimeType="text" sofaString="Joe Doe came
    to the causality complaining of right upper quadrant abdominal
    pain for the past x days. The pain was dull in character. It was
    of sudden onset, did not radiate, and was constant. It did not
    stop from being able to perform her daily activities. She says
    she tried to alleviate the pain taking paracetamol. On
    palpation, her trachea was central, she had normal chest
    explanation with both sides moving symmetrically, apex beat
    couldn't be felt even when asking patient to turn over onto her
    left side. It was associated with nausea and committing and loss
    of appetite. She also said that she would get breathless after
    walking short distances but had no palpitations or chest pain.
    Diagnosed with Chlamydial pneumonia.&#10;" />
  ...
  <org.apache.ctakes.assertion.medfacts.types.Concept _indexed="1"
    _id="10040" _ref_sofa="3" begin="65" end="74" conceptType="PROBLEM"
    conceptText="Chlamydial pneumonia" externalId="0"
    originalEntityExternalId="5532" />
  ...
  <org.apache.ctakes.typesystem.type.refsem.UmlsConcept _id="5514"
    codingScheme="SNOMED" code="233609002" oid="233609002#SNOMED"
    cui="C0000726" tui="T029" />
  ...
</CAS>
```

Pode ser visto que, o texto médico classificado possui uma entrada *medfact* identificada como PROBLEM, e essa entrada possui ainda o código 233609002, referente a sua identificação na terminologia SNOMED CT. Sabendo que o paciente, Joe Doe, foi identificado com esta condição patológica, um módulo Java vai realizar a inserção desta informação na ontologia CPR, como mostra o trecho de código abaixo.

```
<ClassAssertion>
```



```

        <Class IRI="#patient"/>
        <NamedIndividual IRI="#1418639989950_Joe_Doe"/>
    </ClassAssertion>
    <Declaration>
        <NamedIndividual IRI="#CHLAMYDIAL_PNEUMONIA"/>
    </Declaration>
    <ClassAssertion>
        <Class
            IRI="http://purl.org/biotop/biotop.owl#PathologicalCondition"/>
        <NamedIndividual IRI="#1418639989950_233609002"/>
    </ClassAssertion>
    <ObjectPropertyAssertion>
        <ObjectProperty
            IRI="http://purl.org/biotop/1.0/biotop.owl#locatedIn"/>
        <NamedIndividual IRI="#1418639989950_233609002"/>
        <NamedIndividual IRI="#1418639989950_Joe_Doe"/>
    </ObjectPropertyAssertion>
    <ObjectPropertyAssertion>
        <ObjectProperty IRI="http://purl.obolibrary.org/obo/doid#is_a"/>
        <NamedIndividual IRI="#CHLAMYDIAL_PNEUMONIA"/>
        <NamedIndividual IRI="#1418639989950_233609002"/>
    </ObjectPropertyAssertion>
    <DataPropertyAssertion>
        <DataProperty
            IRI="http://www.w3.org/2004/02/skos/core#definition"/>
        <NamedIndividual IRI="#1418639989950_233609002"/>
        <Literal datatypeIRI="#xsd:integer">Chlamydial Pneumonia</Literal>
    </DataPropertyAssertion>

```

É possível ver acima que, foram criadas as instâncias para:

- *patient* "1418639989950\_Joe\_Doe"
- *PathologicalCondition* "1418639989950\_233609002"
- *Disease* "CHLAMYDIAL\_PNEUMONIA".

Ainda, foram criadas as propriedades para associar estas classes, deste modo se tem que "CHLAMYDIAL\_PNEUMONIA" é uma condição patológica localizada no paciente "1418639989950\_Joe\_Doe". Por fim, foi atribuída à condição patológica uma definição, sendo assim ela definida como "Chlamydial Pneumonia".

## 4 Considerações e Trabalhos Futuros

Apesar da efetividade do cTakes e do extenso domínio de conhecimento representado nas ontologias utilizadas, alguns pontos foram identificados como passíveis de melhora em futuros trabalhos.

Como grande parte dos registos clínicos utilizados neste trabalho terem em seus textos clínicos uma descrição do paciente mais voltada para a identificação de sintomas, muitos dos dados obtidos não foram possíveis de serem relacionados

com a *Disease Ontology*, desta forma, a única ontologia a representar o conhecimento clínico destes termos é a CPR. Como exemplo disso, pode ser dado o sintoma *pain*. Apesar de identificado no cTakes e de possuir uma variedade de códigos SNOMED, este sintoma não apresenta nenhuma classe para representá-lo na DO. Isto ocorre em diversos outros sintomas que apesar de serem representados como condições patológicas na CPR não são considerados como doenças na DO.

As bases de dados utilizadas possuem formas de manter o histórico de doenças dos pacientes, mas, nesses casos, são tabelas que armazenam o nome da doença e seus códigos em terminologias como SNOMED CT e ICD9. Assim, não seria necessária uma etapa de processamento de língua natural, pois esta informação já está pronta para ser utilizada sem qualquer tratamento prévio. Como o objetivo deste trabalho foi, buscar em narrativas clínicas em texto livre por termos médicos, esses dados de obtenção direta foram ignorados.

Como uma forma de melhora em trabalhos futuros, pretende-se considerar estes dados de doenças que não necessitam de uma etapa de processamento de língua natural, pois, em muitos dos casos, apesar de o paciente ter uma doença associada em seu histórico clínico, esta não foi informada nas narrativas livres sobre os encontros, tendo sido descritas apenas as ações médicas e os sintomas do paciente. Isto vai permitir um enriquecimento na informação médica deste paciente, pois será possível ter acesso a uma maior variedade de informação a seu respeito.

Ainda será estudada a utilização de uma outra ontologia, especificamente a do SNOMED CT, pois com ela será possível representar o conhecimento também dos sintomas encontrados através da análise dos relatórios médicos, sendo possível, desta forma, manter o conhecimento tanto das doenças como dos sintomas apresentados pelo paciente.

## 5 Conclusão

Os sistemas de recuperação de informação clínica abrangem várias áreas de conhecimento e recorrem a diversas valências para a sua implementação. As tecnologias relativas à representação do conhecimento e raciocínio são consideradas basilares para tais sistemas. A Web Semântica, ao permitir raciocinar e inferir sobre a informação armazenada na World Wide Web, é um formalismo bastante promissor para a elaboração de sistemas de extração de informação.

A representação do conhecimento é um dos principais problemas que devem ser atendidos neste trabalho para executar uma verificação dos pressupostos para a identificação e raciocínio sobre o quadro clínico do paciente. Por isso, deve ser definida uma ontologia para a doença. A representação correta dessa base de conhecimento permite a geração de conhecimento (neste caso, ontologias), que pode ser definido de forma adequada tanto para os termos médicos do domínio quanto para as relações entre esses termos. Além disso, a utilização de terminologias como o SNOMED, permite a criação de um ambiente de interoperabilidade, sendo ainda possível atender a sistemas heterogêneos através de

normas e padrões de integração de sistemas de informação hospitalar, tais como HL7, CEN13606 e openEHR.

Este trabalho apresentou resultados satisfatórios, uma vez que foi possível identificar sintomas e doenças em narrativas clínica de pacientes. Essas informações foram colectadas em bases de dados heterogêneas e apesar dessa característica, foi possível integrar esta informação, abrangendo assim o conhecimento sobre o quadro clínico dos pacientes.

## Referências

1. D. Blumenthal and M. Tavenner. The “meaningful use” regulation for electronic health records. *New England Journal of Medicine*, 363(6):501–504, 2010.
2. M. Corp. Mirth connect introduction and tutorial.
3. T. Cyr, A. Agarwal, and B. Furht. Brief overview of various healthcare tools, methods, framework and standards. In *Handbook of Medical and Healthcare Technologies*, pages 285–295. Springer, 2013.
4. J. Davies, D. Fensel, and F. Van Harmelen. Towards the semantic web. *Ontology-Driven Knowledge Management*. Chichester, 2003.
5. M. Dean, G. Schreiber, S. Bechhofer, F. Van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. Owl web ontology language reference. w3c recommendation, 2004. URL: <http://www.w3.org/TR/owl-ref>, 2006.
6. R. H. Dolin, L. Alschuler, C. Beebe, P. V. Biron, S. L. Boyer, D. Essin, E. Kimber, T. Lincoln, and J. E. Mattison. The hl7 clinical document architecture. *Journal of the American Medical Informatics Association*, 8(6):552–569, 2001.
7. T. J. Eggebraaten, J. W. Tenner, and J. C. Dubbels. A health-care data model based on the hl7 reference information model. *IBM Systems Journal*, 46(1):5–18, 2007.
8. P. L. Elkin, S. H. Brown, C. S. Husser, B. A. Bauer, D. Wahner-Roedler, S. T. Rosenbloom, and T. Speroff. Evaluation of the content coverage of snomed ct: ability of snomed clinical terms to represent clinical problem lists. In *Mayo Clinic Proceedings*, volume 81, pages 741–748. Elsevier, 2006.
9. D. Fensel, F. Van Harmelen, I. Horrocks, D. L. McGuinness, and P. F. Patel-Schneider. Oil: An ontology infrastructure for the semantic web. *IEEE intelligent systems*, 16(2):38–45, 2002.
10. T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.
11. R. V. Guha and D. Lenat. Cyc: A midterm report. *AI magazine*, 11(3):32, 1990.
12. E. Hebel, B. Middleton, M. Shubina, and A. Turchin. Bridging the chasm: effect of health information exchange on volume of laboratory testing. *Archives of internal medicine*, 172(6):517–519, 2012.
13. D. Kalra. Electronic health record standards. 2006.
14. H. Kosorus, J. Honigl, and J. Kung. Using r, weka and rapidminer in time series analysis of sensor data for structural health monitoring. In *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on*, pages 306–310. IEEE, 2011.
15. Y. Liu, L. W.-C. Chan, C.-R. Shyu, and Y. Liu. Editorial for the special issue of knowledge discovery and management in biomedical information systems. *Information Systems Frontiers*, 11(4):345–347, 2009.

16. K. Mahesh. *Ontology development for machine translation: Ideology and methodology*. Citeseer, 1996.
17. J. Potoniec and A. Lawrynowicz. Rmonto: ontological extension to rapidminer. *Poster and Demo Session of the ISWC*, 2011.
18. J. C. C. Robles, A. Domingo, J. M. Colomé, and S. Estevez. Hl7 in personal health system component's integration for mental health treatment. *Journal of Health Informatics*, 3, 2011.
19. M. Saeed, C. Lieu, G. Raber, and R. Mark. Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In *Computers in Cardiology, 2002*, pages 641–644. IEEE, 2002.
20. G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.
21. B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.
22. W. W. Stead, R. B. Haynes, S. Fuller, C. P. Friedman, L. E. Travis, J. R. Beck, C. H. Fenichel, B. Chandrasekaran, B. G. Buchanan, E. E. Abola, et al. Designing medical informatics research and library-resource projects to increase what is learned. *Journal of the American Medical Informatics Association*, 1(1):28–33, 1994.
23. G. Steve, A. Gangemi, and D. M. Pisanelli. Integrating medical terminologies with onlons methodology. *Information Modelling and Knowledge Bases IX*, 9:1, 1998a.
24. J. R. Vest and L. D. Gamm. Health information exchange: persistent challenges and new strategies. *Journal of the American Medical Informatics Association*, 17(3):288–294, 2010.
25. C. Williams, F. Mostashari, K. Mertz, E. Hogin, and P. Atwal. From the office of the national coordinator: the strategy for advancing the exchange of health information. *Health affairs*, 31(3):527–536, 2012.
26. B. A. Wolfe, B. W. Mamlin, P. G. Biondich, H. S. Fraser, D. Jazayeri, C. Allen, J. Miranda, and W. M. Tierney. The openmrs system: collaborating toward an open source emr for developing countries. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1146. American Medical Informatics Association, 2006.